

DOCUMENT RESUME

ED 274 697

TM 860 565

AUTHOR Rowley, Glenn
TITLE Program Evaluation within a Correctional Institution: Threats to the Validity of Evaluative Conclusions.
PUB DATE Apr 86
NOTE 19p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Case Studies; Correctional Institutions; *Correctional Rehabilitation; *Evaluation Methods; *Evaluation Problems; Females; Foreign Countries; *Program Evaluation; Rehabilitation Programs; Research Methodology; Research Problems; *Validity
IDENTIFIERS Australia

ABSTRACT

A year-long case study of the effects of an educational and vocational rehabilitation program in a youth training center was discussed in terms of fourteen issues which might threaten the validity of an evaluation or research project. This correctional institution housed women aged 14 to 21 who were long-term wards of the state, held on remand, or sentenced. The evaluator, who taught a class within the institution, conducted a formative evaluation using observation, interviews, casual conversations, and written records. The evaluation was concerned with both internal validity (program effectiveness for these women and this institution) and external validity (generalizability regarding the funding of similar programs). Threats to the validity of conclusions made about a program were described: (1) historical events previous to the program which affected results; (2) participants' maturation; (3) effects of repeated testing; (4) changes in tests or evaluator's attitudes; (5) statistical regression; (6) attrition; (7) small sample size; (8) bias in selection of sample; (9) obtrusiveness of the inquiry; (10) participant anxiety; (11) diffusion of different treatments; (12) construct validity of causes and effects; (13) communication in problems; and (14) reliance on hearsay evidence rather than direct observation. (GDC)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

PROGRAM EVALUATION WITHIN A CORRECTIONAL INSTITUTION:

THREATS TO THE VALIDITY OF EVALUATIVE CONCLUSIONS

Glenn Rowley

Monash University, Australia

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as
received from the person or organization
originating it.

☐ Minor changes have been made to improve
reproduction quality.

• Points of view or opinions stated in this docu-
ment do not necessarily represent official
OERI position or policy.

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

G. Rowley

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

Paper presented to the Annual Meeting of the American Educational
Research Association, San Francisco, April, 1986.

PROGRAM EVALUATION WITHIN A CORRECTIONAL INSTITUTION:
THREATS TO THE VALIDITY OF EVALUATIVE CONCLUSIONS

Glenn Rowley

Monash University

ABSTRACT

For more than twenty years, researchers using experimental and quasi-experimental methods have looked to the work of Campbell and Stanley (later, Cook and Campbell) for guidance and structure, and for criteria for distinguishing strong from weak design. The contention made in this paper is that these threats to validity should be matters of concern generally in areas of social enquiry; they do not disappear once the researcher chooses to use methods of enquiry other than experimental. Using examples drawn from a year-long case study by the author of the effects of introducing an educational/vocational access programme to a youth training centre for young women, it is demonstrated that the work of Campbell and Stanley has relevance far beyond the experimental/quasi-experimental context in which it was developed and is commonly applied. Threats to the validity of the author's evaluative conclusions are identified and located within these frameworks. Some guidelines are provided for applying these procedures to case study evaluations.

Experimental researchers have long turned to the work of Donald Campbell and his associates for caveats, cautions and frameworks of great assistance in the design and evaluation of their research. In more recent times, with the upsurge in emphasis on the use of qualitative methods in educational evaluations, greater attention has been paid to ways of assessing the validity of qualitative data and the conclusions derived therefrom (e.g., Kennedy, 1984; LeCompte & Goetz, 1982).

Campbell's initial ideas had been widely circulated before they were published as a journal article in 1957; in 1963 this was expanded into a lengthy chapter by Campbell and Stanley in the AERA Handbook of Research on Teaching (1963); republished as an 84-page monograph (Campbell and Stanley, 1966); and finally expanded into a 405-page text by Cook and Campbell (1979). During these times, the ideas have been revised, rethought and updated many times (e.g., Winch and Campbell, 1979; Campbell, 1975). The ten threats to validity proposed by Campbell (1957) were expanded to 12 by Campbell and Stanley (1963), to 15 by Campbell (1969) and Winch & Campbell (1970) and to 32 by Cook and Campbell (1979). Initially, Campbell's focus was an internal validity; Campbell and Stanley (1963) extended their treatment to include external validity, and Cook and Campbell (1979) added statistical conclusion validity and construct validity as further considerations.

Because of the context in which these ideas have been presented, it has been common to think of them as having application to experimental research only. Case study researchers, for example, have typically not regarded these considerations as relevant to their work. Indeed, Campbell and Stanley's initial comments on "the one-shot case study":

... such studies have such a total absence of control as to be of almost no scientific value. The design is introduced here as a minimum reference point. ... It seems well-nigh unethical at the present time to allow, as these or dissertations in education, case studies of this nature (i.e., involving a single group observed at one time only) (Campbell and Stanley, 1966, pp. 6-7).

have often been perceived as disparaging to all case study research, rather than to the particular type of case study ("one shot") referred to, which has more in common with Rist's "Blitzkreig ethnography" (1980) than to genuine detailed case studies conducted over adequate periods of time. Recently LeCompte and Goetz (1982) structured a discussion of

reliability and validity in ethnographic research around the concepts of internal and external reliability and validity. While the terms were borrowed from the Campbell-Stanley framework, the meanings attached were quite different.

The author's recent experience with a study which necessitated the use of case study methods has served to convince him that there is much of value to the case study researcher in these writings. The major purpose of this paper is to illustrate the variety of threats to the validity of evaluation-based conclusions by submitting his own study to the kind of scrutiny to which experimental studies are routinely subjected.

The study

Recently the author was responsible for a year-long study of the effects of introducing an educational/vocational access programme to a youth training centre for young women. The programme was directed towards those young women who, through repeated offences and/or the length of their sentences, had become long-term detainees. The aim of the programme was, through the provision of appropriate educational and vocational counselling, to encourage these young women to consider a wider range of options in life, and to help them to develop modes of living which were both personally satisfying and within the law.

The institution can accommodate up to 95 young women, each of which would be either:

1. a ward of state
2. held on remand, or
3. sentenced.

Only a small proportion of admittances are on sentence, the majority being returned wards, either on Care and Protection Applications or transferred from other institutions, often because they present severe management problems. Their ages range from 12 to 21 years, although the admission of youth younger than 14 or older than 18 is relatively uncommon.

The Educational/Vocational Access Programme, which is the focus of this report, was formulated specifically to meet the needs of older youth, those on sentence, and those who have a history of returning to the

institution. These young women are housed largely in one section of the institution, and this section was the base for the operation of the programme. It should be noted, though, that young women from other sections were at times involved in the programme, either by referral, or by commencing their involvement while in the appropriate section, and continuing it after moving to another.

The young women involved in this programme were aptly described in the original grant submission in these terms:

... a core group of young women who continue to come back to the institution either as wards of the state or on a sentence. The options which have been provided for them up till this time have been inappropriate. It is the needs of this group of young women that this programme seeks to address.

Of this core group, youth tend to:

1. be early school leavers,
2. have had a history of school failure,
3. be thoroughly disenchanted with previous school experience,
4. have had little or no meaningful employment - employment obtained falls within the unskilled area and may have been exploitative
5. be recidivists over a long period of time,
6. go straight back to the drug scene, or drift there on release,
7. believe that the 'straight' world has little to offer them - life is short ... enjoy it while you can.

In summary these youth have a history of failure in the transition from the institution to the community despite the many plans made for them. Many drift to St. Kilda, prostitution, the drug scene and crime. The 'scene' offers them the feeling of excitement, acceptance, belonging and purpose.

The aims of the Educational/Vocational Access Programme, as stated in the application, were:

1. to provide information regarding education and employment so that young women can make informed decisions regarding their future;

2. to encourage young women to investigate long-term career prospects,
3. to explain labour market trends and their implications,
4. to provide linkages with outside educational institutions, e.g. TAFE¹, to (a) establish access to appropriate courses, and (b) develop module courses to be run at the institution,
5. to acquaint young women with their rights and responsibilities within the work force and the community ... to encourage them to exercise these rights and responsibilities,
6. to develop alternative programmes for employment opportunities,
7. to provide ongoing support for young women in their educational and employment pursuits when they leave the institution.

In designing and implementing the evaluation, I was given a relatively free hand. It was clear that success in achieving the aims listed above should be one focus of the evaluation, but beyond that I was subject to no restrictions nor prescriptions. The report focussed on outcomes at the individual level (the effects of the access programme on particular young women), and at the institutional level (the effects on the institution). In addition it tried to present a comprehensive portrait of the programme in operation and to be formative rather than just summative. The results of the study have been reported in a Report to the Victorian Ministry of Employment and Training (Rowley, 1985). This paper reflects on these findings, using a framework based on the original by Campbell and Stanley (1963), but expanded in the manner of Cook and Campbell (1979) and Kennedy (1984).

Evaluation strategy

The opportunity to teach a regular class within the institution afforded me a role which was non-threatening to participants in the programme, a reason for being regularly present in the institution, and regular personal contact with the young women who were the focus of the access programme. Thus data for the evaluation came from my own observations over a long period of time, from casual conversation with staff and young women, from more structured interviews with participants, and from written records. Further details are spelled out in the Report (Rowley, 1985).

Internal and external validity

A distinction fundamental to experimental research is that between internal and external validity. As originally described by Campbell and Stanley (1963).

Internal validity is the basic minimum without which any experiment is uninterpretable: Did in fact the experimental treatments make a difference in this specific experimental instance? External validity asks the question of generalizability: To what populations, settings, treatment variables, and measurement variables can this effect be generalized? (p. 175).

Further, they noted that:

While internal validity is the sine qua non, and while the question of external validity, like the question of inductive inference, is never completely answerable, the selection of designs strong in both types of validity is obviously our ideal. This is particularly the case for research on teaching, in which generalization to applied settings of known character is the desideratum (p. 175).

The study which formed the focus of this paper was not an experiment, either in design or in execution. The researcher went into a situation in which an innovative programme was already functioning, charged with the task of evaluating what was happening there. The sponsor of the research (the Ministry of Employment and Training) sought conclusions about the programme at two levels:

1. What are the effects of this programme with respect to the particular young women to whom it was directed, and on the institution within which the programme was housed?
2. What recommendations can be provided to the Ministry with respect to the funding of this programme, and to the general question of future funding of programmes in situations like this one?

Although it was not spelled out in the language of Campbell and Stanley it was clear to see that the Ministry was specifically concerned about questions of internal and external validity.

For work like this, best described as evaluation by case study, my basic aim was to come to an understanding of what was happening in the institution, and what elements of that might reasonably be attributed to the presence of an innovative programme within the institution. Much case study work is of this nature, although in many cases the focus would be less specific and less outcome-oriented than it was in this case. Nevertheless, conclusions are reached as to what is happening and why, and how events are viewed by participants. A key question (perhaps the most important one) in judging the quality of the research is 'what steps has the researcher taken to protect himself or herself from being misled, either by events or by people's accounts of events?' Experimental researchers use specific means to do this - the science of experimental design (control groups, random assignment, placebos, blind and double-blind designs, factorial designs, Latin square designs, etc.) is designed for this purpose. Case study research is typically and desirably less intrusive, but the concern for accuracy of interpretation of events remains. Thus case study researchers seek corroboration both of events and of their interpretation of events; they seek to triangulate, they seek the help of "critical friends", and so on. The methods are different, the control is not as logically "tight", but if well done, their methods can yield acceptable levels of protection against known sources of error.

External validity has typically been a major preoccupation for survey researchers, whose findings would be of no interest if they could not be generalized to a wider population than the group from which they were obtained. Neither experimental nor case study research has typically used strong procedures for ensuring generalizability, as does the survey researcher in drawing a carefully representative sample. For both case studies and experiments, generalizability is claimed on the basis of an accumulation of findings in a variety of settings. The basis for generalization is generally intuitive, although meta-analysis (Glass, 1976) provides a means of generalizing which relies less on intuition and more on rules of evidence. Its use is, however, restricted to situations in which findings can be expressed in quantitative terms.

Threats to validity

The greatest contribution of the work of Campbell and his collaborators has been in outlining the threats to validity of which

experimental researchers must be aware, and against which their research designs should afford protection. It is the contention of this paper that the questions are as relevant to case study research as they are to experimental research. To this end, the remainder of the paper outlines several of the threats noted by Campbell and Stanley (1963), Winch & Campbell (1969), Cook & Campbell (1979), and considers their applicability to this particular example of case study research.

1. "History" refers to events during the course of an experiment, the effects of which might mistakenly be interpreted as outcomes of the experimental treatment. An example cited by Campbell and Stanley (1966) is a study on the effects of Nazi propaganda carried out during 1940, during which France fell. Events of that magnitude would hardly not be taken account of, but less notable events whose effects might not be known must always be a concern. In experimental studies, random allocation into treatment groups is a major safeguard against this threat to validity. This does not, however, provide any protection against events taking place within the classroom or setting of one group but not the other, e.g., physical mishaps, isolated cases of misbehaviour, etc.

Any event in the life of one of the young women involved in the study is potentially, at least, a threat to validity in this sense. Many events which took place during the year quite independently of the programme could have seriously jeopardized the outcomes that it was striving for - in particular, absconsions, the arrival of particular inmates holding extreme values or attitudes, and incidents outside in which our young women unwittingly became involved. But most worrisome is the class of events which might conceivably affect the young women as a group - among these one would certainly have to include changes in the economy, and in particular, in rates of unemployment. When unemployment rates among young women are around 20 percent, minor changes in overall rates can have the most dramatic effects on the employment prospects of this most vulnerable of groups. To focus the study entirely or even largely on the outcome of gaining a job would leave the study open to this threat. Changes in rules concerning eligibility for unemployment benefits, for student allowances, all of which failed to occur during the time of the study, would have been serious threats had they taken place, since they would certainly have had dramatic and systematic effects on the young women's prospects.

2. Maturation refers to long-term systematic changes which could affect the outcomes of an experimental study - in Campbell and Stanley's words, "all of those biological or psychological processes which systematically vary with the passage of time, independent of specific external events (1963, pp. 177-178)." They cite specific examples - growing older, hungrier, more tired, more bored, etc., and the medical phenomenon of "spontaneous remission", which is not spontaneous in the usual sense, but merely represents the cumulation of many possibly long-term processes and events. Not specifically mentioned, but ever present in studies with young people which extend over a period of time, is the risk of misinterpreting differing rates of development between girls and boys.

The young women involved in this programme were all in their mid to late teens, a time in which many people go through substantial and sometimes unexpected changes - physically, emotionally and intellectually. In addition, many of them had been through the most traumatic experiences, deprived of love and security, and abused by others and themselves. Institutional life had provided them with a measure of security which many had not known before, and with it a lack of freedom which many resented bitterly. At the time in which they were involved in the programme most of the young women were nearing the end of this phase of their lives. Ahead of them lay what? Adult prison? A life on the fringes of the law? A "straight" life? Clearly, they were at or approaching a time in their lives where important decisions were going to be made, and when these decisions were made their lives would undergo big changes. But to attribute these changes to the access programme without further examination would be hazardous indeed.

3. Previous testing may increase a person's ability to perform on subsequent tests, and the improvement might wrongly be attributed to an experimental intervention. Equally, repeated testing or observation may result in boredom or lack of motivation, and lower scores may be used wrongly to discredit an experimental intervention.

My presence as an observer and the way in which I was perceived by the young women may have influenced the way they responded to me and the information that I received. Features of the study that were intended to minimize this problem were my presence over a long time period and my "legitimate" role as one of their teachers. It would be presumptuous for

me to claim that my observations and interviews were nonreactive; at best it can be said that steps were taken to minimize the reactivity of the process.

4. Instrumentation refers to changes in the functioning or measuring properties of a test during the course of an experiment. Two examples: first, a test may measure a different skill when used a second time - e.g., a pretest which involves a high level of reasoning skill may merely measure recall when used as a posttest, particularly if the instruction is closely related to the test content. Second, the measurement properties of the test may be quite different due to scores falling in a different part of the range - e.g., ceiling effects.

In this case the instrument was the experimenter, the data were my own observations, records of outcomes, and the comments and observations of participants and other staff. In the early part of the year I was acutely conscious of my inadequacy as an instrument of observation. The complexities of life in a correctional institution are considerable, and not easily understood by a newcomer. Consequently initial accounts of events were contradictory and confusing, and it was only with the benefit of hindsight that they began to make sense. My mood and my attitude toward the programme fluctuated according to events and to which groups or persons I was most closely associating with at the time. As I spoke with supporters of the programme I became a supporter; as I spoke with critics I became a critic. Admiration and exasperation succeeded one another like day and night. How this affected the information which was provided to me and my own ability to interpret it is difficult to judge. By the end of the year I felt more confident in my observations, partly because I understood the context so much more, and partly because large segments of what was to become my report had already been read and reacted to by participants.

5. Statistical regression. Whenever individuals are selected because of their extreme scores on any measure, the probability is that, on retesting, their scores will be less extreme. High scorers will come down, low scorers will improve, both groups will regress towards the mean. So it is that poor readers selected for a remedial programme on the basis of their low test scores are most likely to improve those scores on retesting - regardless of the merits of the programme. This does not mean that their reading will improve; just that whatever chance factors

had acted to depress their scores on the initial testing are unlikely to act uniformly in the same direction again on the retest. As long as chance plays a part, statistical regression will occur. When your luck is at its worst, it is most likely to improve.

This study was not based on statistical evidence, but the same principle applies. For many (but not all) of the young women, their confinement to an institution represents a low point in their lives; a time when the odds favouring an improvement in their situation are probably as favourable as they are ever going to be. Such improvement, when it occurs, would tend to cast a favourable but undeserved glow on whatever programmes were in operation at the time. This problem for the researcher is compounded by a non-chance factor; as the young women approach 18 years of age and the prospect of adult prison, their incentive to improve their situations may be greater than it has been in the past.

6. Experimental mortality refers to the particular problem faced in experimental work where there is an element of voluntarism about continued participation in the project. Students drop out of classes, clients drop out of therapy, etc. If the nature of the treatment is such that it contributes to differential rates of dropping out between experimental and control group, there will be concern that, not matter how carefully selected and equated the groups may have been when the experiment began, the groups remaining at the end of the experiment could be far from equivalent. Mortality can also constitute a threat to external validity; even if experimental and control groups are equally affected, generalization is now possible only to the population of "survivors," which may be quite different to the population from which the groups were originally selected, and for which the experimental treatment was intended.

More than sixty young women involved in the programme over the year, but typically no more than ten at any one time within the institution. Participation was always voluntary, although refusals to participate and/or dropping out by girls within the institution were never serious problems. But dropping out, in the sense of leaving the institution was always a disruptive factor. Every week young women were released; others were admitted (or re-admitted). Absconding was not uncommon. Whenever one young women was prepared to attempt an escape, there would be others willing to go with her. For young women ^{on} sentence, it was possible to

plan a programme with a particular release date in mind. For those under Care and Protection orders, the date of release was not so predictable, and was likely to be advanced by a young woman gaining an offer of employment - a factor which made long-term planning difficult. But from the point of view of an evaluator, it meant that the young women for whom the programme had produced short-term successes often did not remain in the programme.

7. Instability was added to the original Campbell and Stanley list by Winch and Campbell (1969), and greatly elaborated by Cook and Campbell (1979). It refers to the insensitivity of many studies in detecting treatment effects because often differences between treatment groups can be swamped by the natural variation which occurs within any treatment groups. With small samples the problem is most acute; greater sensitivity comes with larger sample sizes. In any case, the widespread use of statistical significance tests enables the researcher to estimate the seriousness of the problem, and if not to eliminate the threat, at least to confine it within preset probabilistic limites.

The programme dealt with a limited number of young women and at the end of one year, clear outcomes had been achieved for very few of them. A few outstanding successes, a few failures, and the overall impression might have been quite different. Although all relevant staff were interviewed, they too were limited in number, and the impressions of one more person, whether a supporter or a critic can have a significant impact. In both cases, conclusions were based upon a preponderance of evidence, rather than on statistical tests. But human variation is so great, both amongst the subjects of the programme and among participants and observers, that one needs to consider the potential instability in one's impressions due to the limited number of informants.

8. Selection. In a "true" experiment, random selection is used to see that experimental and control groups are equivalent. This does not mean that they do not differ from one another, but that pre-existing differences between comparison groups occur only through the operation of chance. The probability of differences of any specified size occurring by chance can be determined. In situations which do not allow random allocation of subjects into comparison groups, the possibility of

systematic differences between them cannot be discounted. In these circumstances, it is said that selection is a threat to the internal validity of the research.

Although this was not an experimental study, the implied question at the back of everybody's minds was "how different would it be if the access programme were not operating?" No formal comparisons were made, but the implied comparison was ever present. Hard evidence about the comparison was not obtained, indeed comparison groups were never formally identified. In the minds of many of those who spoke to me about the programme was the implied comparison with previous years, but differences in intake and in economic circumstances pose huge threats to the validity of any such comparisons.

9. Obtrusiveness of the inquiry (Kennedy, 1984). This includes the Campbell and Stanley notions of "artificiality" and "reactive effects of testing." Awareness by the participants that they are taking part in an experiment or an evaluation, and the artificiality of many experimental arrangements, cast doubt on the extent to which the results of many studies can be generalized. Frequently there is concern in experimental studies that pretests, interviews or observations might sensitize the subjects of an experiment in a way which would make the experimental treatment appear more effective. In case-study evaluations, there must be concern that social desirability affects subjects' responses sufficiently to invalidate some conclusions.

Although in case study work generally, it is possible that the presence of an observer may change the situation observed so much that conclusions cannot legitimately be drawn about what would have happened in the absence of the observer. In this particular study I do not think that my presence had any important effect on the young women and their reaction to the access programme, although there were indications of a reluctance to provide me with information which might be seen as damaging to the programme. One particular incident early in the year illustrated the need for caution. My first interview with a group of young women in the programme was followed by a report back to the project officers that they had better watch out for me; I was going around asking questions about them. The young women accepted with some incredulity the project officers' assurances that it was quite OK, and I was doing it with their approval.

10. Evaluation apprehension (Cook & Campbell, 1979) refers to the effects on respondents of anxiety about their being evaluated. This may cause respondents to strive to present themselves in a more favourable light, and if they have a personal stake in the programme, to present the programme in a more favourable light.

Project officers, those who initiated the access project, and those who supported it must have felt a degree of apprehension about the evaluation, although to their credit it was never obvious in their relationships with me. But it is scarcely possible for project officers in particular not to equate an evaluation of their project with an evaluation of themselves. Apprehension can only have been made more acute by the short-term nature of their employment and the clear implication that decisions about future funding would take into account the findings of the evaluation. The extent to which this affected their work is impossible to judge: the fear that it might have limits generalization to other situations.

11. Imitation or diffusion of treatments (Cook & Campbell, 1979) refers to the possibility that comparisons between experimental and control groups can be invalidated if there is sufficient communication between the groups to enable aspects of the treatment given to the experimental group to be passed on to the control group. This is a particular problem in research within schools, where the matching of classes within schools would otherwise be a strong design feature.

It had soon become apparent that the access programme was viewed as threatening by staff associated with another programme already established within the institution. A degree of cooperation was established between the two programmes, young women passed between them, and in one particular case, both programmes claimed credit for the successes gained by the same young woman. Whilst every example of cooperation between the two programmes was naturally to be applauded, it also presented problems for the evaluator: to what should the success be attributed? In fact the level of cooperation between the two programmes came to be seen as one of the positive outcomes that flowed from the introduction of the access programme.

12. Construct validity of causes and effects (Cook & Campbell, 1979) refers to the concern that both causes and effects may be inadequately defined and represented. What are the essential features of a successful treatment, and what may be disregarded? How can we be sure that all of the important effects of a programme have been observed? Bracht & Glass (1968) noted that generalization is desired not just across persons, but also across treatments, settings, times and outcomes. Unless treatments, settings and outcomes are well-defined, the degree of generalization possible will be uncertain.

It was clear to me the Ministry wanted guidance on the desirability of funding other similar projects in other institutions. Clearly, they wanted generalizability across settings, but I could provide no evidence for it. The nature of the access programme was not comprehensively documented, and it was only by observing it in action that I began to develop an understanding of the nature of the "treatment". Even that was elusive, as the programme changed quite dramatically during the course of the year as staff responded to events. So, even if the programme is judged to be a roaring success, how confident can we be that another programme with a similar title will be as successful in another setting? Not very, but this is the type of generalization that was sought.

13. Problems of communication. Kennedy (1984) has noted the extent to which the inherent ambiguity of the language used by the participants can be a threat to the validity of naturalistic data. The greater the initial distance between the case study worker and the participants, and the shorter the time scale of the research, the more likely it is that they will use and understand words differently.

The fact that the study extended over a whole year was a point in its favour. This enabled me time to develop an understanding of the way in which people described their situations. But it was time that was desperately needed. For one unfamiliar with correctional institutions, there is a wealth of terminology, jargon and "slang" to be learned. And some of the language used by the young women, needless to say, was a little different to that normally used by my colleagues at Monash.

14. Inability of the evaluator to observe directly. As Kennedy (1984) pointed out, much of the data upon which conclusions were based were in fact, verbal testimony about behavior rather than direct observations of behavior. Necessarily, some of the evidence provided to me was hearsay evidence.

This was a serious limitation of this study, as it must be of case-study evaluations, generally. We use some of the methods of ethnographers, but we do not do ethnographies. I relied on accounts of events provided to me by participants in the events. Most of the participants had axes of one kind or another to grind, and made no secret of it. My task, like that of any evaluator, was to try to understand how the conflicts in their accounts came about.

I also relied on hearsay evidence with regard to young women who had passed through the programme prior to my arrival, had absconded or otherwise departed during the days of the week that I was absent. Information about the whereabouts and activities of the young women after their release was difficult to obtain and virtually impossible to verify. When I was advised that one or other young woman who had departed was "back on the streets", I accepted the advice without insisting on direct confirmation. Perhaps I was too timid in this regard.

Conclusions

There is much of value for case study researchers/evaluators in the methodological writings on experimental research, and in this paper I have attempted to demonstrate some of it. Drawing upon my own research for potential sources of invalidity has enabled me to be much more critical than would have been possible had I used the research of my colleagues.

Much more can yet be done. Many more of the 33 threats outlined by Cook and Campbell (1979) are relevant to case study methods, and a more exhaustive analysis of these threats in the non-experimental context would be of value. Kennedy (1984) has provided a useful outline of threats to the validity of qualitative data, but further exposition (perhaps within the framework suggested by LeCompte and Goetz, 1982) of threats that are unique to the case study and/or to ethnographic research would be of great value to researchers and to readers of research.

REFERENCES

- Bracht, G.H., & Glass, H.V. (1968). The external validity of experiments. American Educational Research Journal, 5, 437-474.
- Campbell, D.T. (1957). Factors relevant to the validity of experiments in social settings. Psychological Bulletin, 54, 297-312.
- Campbell, D.T. (1969). Reforms as experiments. American Psychologist, 24, 409-429.
- Campbell, D.T. (1975). "Degrees of freedom" and the case study. Comparative Political Studies, September 1975, 178-193.
- Campbell, D.T., & Stanley, J.C. (1963). Experimental and quasi-experimental designs for research on teaching. In N.L. Gage (Ed.), Handbook of Research on Teaching, Chicago: Rand-McNally. (also published as Experimental and Quasi-experimental Designs for Research. Chicago: Rand-McNally, 1966).
- Cook, T.D., & Campbell, D.T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Boston: Houghton-Mifflin.
- Glass, G.V. (1976). Primary, secondary and meta-analysis of research. Educational Researcher, 5, 10,3-8.
- Kennedy, M.M. (1984). Assessing the validity of qualitative data. Educational Evaluation and Policy Analysis, 6, 367-377.
- LeCompte, M.D., & Goetz, J.P. (1982). Problems of reliability and validity in ethnographic research, Review of Educational Research, 52, 31-60.
- Rist, R. (1980). Blitzkreig ethnography: on the transformation of a method into a movement. Educational Researcher, 9, 2, 8-10.
- Rowley, G.L. Educational and Vocational Access for Incarcerated Youth. Victoria: Ministry of Employment and Training, 1985.
- Winch, R.F., & Campbell, D.T. (1969). Proof? No! Evidence? Yes! The significance of tests of significance. American Sociologist, 4, 140-143.